

Regularized Least-Mean-Square Algorithms

Yilun Chen, *Student Member, IEEE*, Yuantao Gu, *Member, IEEE*, and Alfred O. Hero, III, *Fellow, IEEE*

Abstract—We consider adaptive system identification problems with convex constraints and propose a family of regularized Least-Mean-Square (LMS) algorithms. We show that with a properly selected regularization parameter the regularized LMS provably dominates its conventional counterpart in terms of mean square deviations. We establish simple and closed-form expressions for choosing this regularization parameter. For identifying an unknown sparse system we propose sparse and group-sparse LMS algorithms, which are special examples of the regularized LMS family. Simulation results demonstrate the advantages of the proposed filters in both convergence rate and steady-state error under sparsity assumptions on the true coefficient vector.

Index Terms—LMS, NLMS, convex regularization, sparse system, group sparsity, l1 norm

I. INTRODUCTION

The Least Mean Square (LMS) algorithm, introduced by Widrow and Hoff [1], is a popular method for adaptive system identification. Its applications include echo cancelation, channel equalization, interference cancelation and so forth. Although there exist algorithms with faster convergence rates such as the Recursive Least Square (RLS) methods, LMS-type methods are popular because of its ease of implementation, low computational costs and robustness.

In many scenarios often prior information about the unknown system is available. One important example is when the impulse response of the unknown system is known to be sparse, containing only a few large coefficients interspersed among many small ones. Exploiting such prior information can improve the filtering performance and has been investigated for several years. Early work includes heuristic online selection of active taps [2]–[4] and sequential partial updating [5], [6]; other algorithms assign proportional step sizes of different taps according to their magnitudes, such as the Proportionate Normalized LMS (PNLMS) and its variations [7], [8].

Motivated by LASSO [9] and recent progress in compressive sensing [10], [11], the authors in [12] introduced an ℓ_1 -type regularization to the LMS framework resulting in two sparse LMS methods called ZA-LMS and RZA-LMS. This methodology was also applied to other adaptive filtering frameworks such as RLS [13], [14] and projection-based adaptive algorithms [15]. Inheriting the advantages of conventional LMS methods such as robustness and low computational costs, the sparse LMS filters were empirically demonstrated

to achieve superior performances in both convergence rate and steady-state behavior, compared to the standard LMS when the system is sparse. However, while the regularization parameter needs to be tuned there is no systematical way to choose the parameter. Furthermore, the analysis of [12] is only based on the ℓ_1 penalty and not applicable to other regularization schemes.

In this paper, we extend the methods presented in [12], [16] to a broad family of regularization penalties and consider LMS and Normalized LMS algorithms (NLMS) [1] under general convex constraints. In addition, we allow the convex constraints to be time-varying. This results in a regularized LMS/NLMS¹ update equation with an additional sub-gradient term. We show that the regularized LMS provably dominates its conventional counterpart if a proper regularization parameter is selected. We also establish a simple and closed-form formula to choose this parameter. For white input signals, the proposed parameter selection guarantees dominance of the regularized LMS over the conventional LMS. Next, we show that the sparse LMS filters in [12], *i.e.*, ZA-LMS and RZA-LMS, can be obtained as special cases of the regularized LMS family introduced here. Furthermore, we consider a group-sparse adaptive FIR filter response that is useful for practical applications [8], [17]. To enforce group sparsity we use $\ell_{1,2}$ type regularization functions [18] in the regularized LMS framework. For sparse and group-sparse LMS methods, we propose alternative closed-form expressions for selecting the regularization parameters. This guarantees provable dominance for both white and correlated input signals. Finally, we demonstrate performance advantages of our proposed sparse and group-sparse LMS filters using numerical simulation. In particular, we show that the regularized LMS method is robust to model mis-specification and outperforms the contemporary projection based methods [15] for equivalent computational cost.

The paper is organized as follows. Section II formulates the problem and introduces the regularized LMS algorithm. In Section III we develop LMS filters for sparse and group-sparse system identification. Section IV provides numerical simulation results and Section V summarizes our principal conclusion. The proofs of theorems are provided in the Appendix.

Notations: In the following parts of paper, matrices and vectors are denoted by boldface upper case letters and boldface lower case letters, respectively; $(\cdot)^T$ denotes the transpose operator, and $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norm of a vector, respectively.

¹We treat NLMS as a special case of the general LMS algorithm and will not distinguish the two unless required for clarity.

Y. Chen and A. O. Hero are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. Tel: 1-734-763-0564. Fax: 1-734-763-8041. Emails: {yilun, hero}@umich.edu.

Y. Gu is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Tel: +86-10-62792782, Fax: +86-10-62770317. Email: gyt@tsinghua.edu.cn.

This work was partially supported by AFOSR, grant number FA9550-06-1-0324.

II. REGULARIZED LMS

A. LMS framework

We begin by briefly reviewing the framework of the LMS filter, which forms the basis of our derivations to follow. Denote the coefficient vector and the input signal vector of the adaptive filter as

$$\hat{\mathbf{w}}_n = [\hat{w}_{n,0}, \hat{w}_{n,1}, \dots, \hat{w}_{n,N-1}]^T \quad (1)$$

and

$$\mathbf{x}_n = [x_n, x_{n-1}, \dots, x_{n-N+1}]^T, \quad (2)$$

respectively, where n is the time index, x_n is the input signal, $\hat{w}_{n,i}$ is the i -th coefficient at time n and N is the length of the filter. The goal of the LMS algorithm is to identify the true system impulse response \mathbf{w} from the input signal x_n and the desired output signal y_n , where

$$y_n = \mathbf{w}^T \mathbf{x}_n + v_n. \quad (3)$$

v_n is the observation noise which is assumed to be independent with x_n .

Let e_n denote the instantaneous error between the filter output $\hat{\mathbf{w}}_n^T \mathbf{x}_n$ and the desired output y_n :

$$e_n = y_n - \hat{\mathbf{w}}_n^T \mathbf{x}_n. \quad (4)$$

In the standard LMS framework, the cost function L_n is defined as the instantaneous square error

$$L_n(\hat{\mathbf{w}}_n) = \frac{1}{2} e_n^2$$

and the filter coefficient vector is updated in a stochastic gradient descent manner:

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n - \mu_n \nabla L_n(\mathbf{w}_n) = \hat{\mathbf{w}}_n + \mu_n e_n \mathbf{x}_n, \quad (5)$$

where μ_n is the step size controlling the convergence and the steady-state behavior of the LMS algorithm. We refer to (5) as the conventional LMS algorithm and emphasize that μ_n can be both time-varying and functions of \mathbf{x}_n . For example,

$$\mu_n = \frac{\alpha_n}{\|\mathbf{x}_n\|_2^2} \quad (6)$$

yields the normalized LMS (NLMS) algorithm with variable step size α_n .

B. Regularized LMS

Conventional LMS algorithms do not impose any model on the true system response \mathbf{w} . However, in practical scenarios often prior knowledge of \mathbf{w} is available. For example, if the system is known to be sparse, the ℓ_1 norm of \mathbf{w} can be upper bounded by some constant [9]. In this work, we study the adaptive system identification problem where the true system is constrained by

$$f_n(\mathbf{w}) \leq \eta_n, \quad (7)$$

where $f_n(\cdot)$ is a convex function and η_n is a constant. We note that the subscript n in $f_n(\cdot)$ allows adaptive constraints that can vary in time. Based on (7) we propose a regularized instantaneous cost function

$$L_n^{\text{reg}}(\hat{\mathbf{w}}_n) = \frac{1}{2} e_n^2 + \gamma_n f_n(\hat{\mathbf{w}}_n) \quad (8)$$

and update the coefficient vector by

$$\begin{aligned} \hat{\mathbf{w}}_{n+1} &= \hat{\mathbf{w}}_n - \mu_n \nabla L_n^{\text{reg}}(\hat{\mathbf{w}}_n) \\ &= \hat{\mathbf{w}}_n + \mu_n e_n \mathbf{x}_n - \rho_n \partial f_n(\hat{\mathbf{w}}_n), \end{aligned} \quad (9)$$

where $\partial f_n(\cdot)$ is the sub-gradient of the convex function $f_n(\cdot)$, γ_n is the regularization parameter and $\rho_n = \gamma_n \mu_n$.

Eq. (9) is the proposed regularized LMS. Compared to its conventional counterpart, the regularization term, $-\rho_n \partial f_n(\hat{\mathbf{w}}_n)$, always promotes the coefficient vector to satisfy the constraint (7). The parameter ρ_n is referred to as the regularization step size. Instead of tuning ρ_n in an *ad hoc* manner, we establish a systematic approach to choosing ρ_n .

Theorem 1. Assume both $\{x_n\}$ and $\{v_n\}$ are Gaussian independent and identically distributed (i.i.d.) processes that are mutually independent. For any $n > 1$

$$E \|\hat{\mathbf{w}}_n - \mathbf{w}\|_2^2 \leq E \|\hat{\mathbf{w}}'_n - \mathbf{w}\|_2^2 \quad (10)$$

if $\hat{\mathbf{w}}_0 = \hat{\mathbf{w}}'_0$ and $\rho_n \in [0, 2\rho_n^*]$, where \mathbf{w} is the true coefficient vector and $\hat{\mathbf{w}}'_n$ and $\hat{\mathbf{w}}_n$ are filter coefficients updated by (5) and (9) with the same step size μ_n , respectively. ρ_n^* is calculated by

$$\rho_n^* = \max \left\{ (1 - \mu_n \sigma_x^2) \frac{f_n(\hat{\mathbf{w}}_n) - \eta_n}{\|\partial f_n(\hat{\mathbf{w}}_n)\|_2^2}, 0 \right\} \quad (11)$$

if μ_n are constant values (LMS), or

$$\rho_n^* = \max \left\{ (1 - \alpha_n/N) \frac{f_n(\hat{\mathbf{w}}_n) - \eta_n}{\|\partial f_n(\hat{\mathbf{w}}_n)\|_2^2}, 0 \right\} \quad (12)$$

if μ_n is chosen using (6) (NLMS), where N is the filter length, σ_x^2 is the variance of $\{x_n\}$ and η_n is an upper bound of $f_n(\mathbf{w})$ defined in (7).

The proof of Theorem 1 is provided in the Appendix.

Remark 1. Theorem 1 shows that with the same initial condition and step size μ_n , the regularized LMS algorithm provably dominates conventional LMS when the input signal is white. The parameter ρ_n^* in (11) or (12) can be used as the value for ρ_n in (9) to guarantee that regularized LMS will have lower MSD than conventional LMS. The value ρ_n^* only requires specification of the noise variance and η_n which upper bounds the true value $f_n(\mathbf{w})$. Simulations in latter sections show that the performance of the regularized LMS is robust to misspecified values of η_n .

Remark 2. Eq. (11) and (12) indicate that to ensure superiority the regularization is only “triggered” if $f_n(\hat{\mathbf{w}}_n) > \eta_n$. When $f_n(\hat{\mathbf{w}}_n) \leq \eta_n$, $\rho_n^* = 0$ and the regularized LMS reduces to the conventional LMS.

Remark 3. The closed form expression for ρ_n^* is derived based on the white input assumption. Simulation results in latter sections show that the (11) and (12) are also empirically good choices even for correlated input signals. Indeed, in the next section we will show that provable dominance can be guaranteed for correlated inputs when the regularization function is suitably selected.

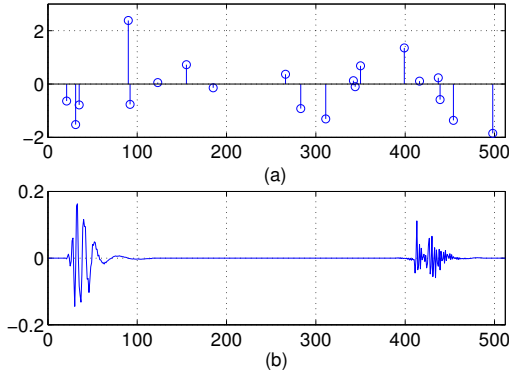


Fig. 1. Examples of (a) a general sparse system and (b) a group-sparse system.

III. SPARSE SYSTEM IDENTIFICATION

A sparse system contains only a few large coefficients interspersed among many negligible ones. Such sparse systems arise in many applications such as digital TV transmission channels [17] and acoustic echo channels [8]. Sparse systems can be further divided into general sparse systems and group-sparse systems, as shown in Fig. 1 (a) and Fig. 1 (b), respectively. Here we apply our regularized LMS to both general and group sparse system identification. We show that ZA-LMS and RZA-LMS in [12] are special examples of regularized LMS. We then propose group-sparse LMS algorithms for identifying group-sparse systems.

A. Sparse LMS

For a general sparse system, the locations of active non-zero coefficients are unknown but one may know an upper bound on their number. Specifically, we will assume that the impulse response \mathbf{w} satisfies

$$\|\mathbf{w}\|_0 \leq k, \quad (13)$$

where $\|\cdot\|_0$ is the ℓ_0 norm denoting the number of non-zero entries of a vector, and k is a known upper bound. As the ℓ_0 norm is non-convex it is not suited to the proposed framework. Following [9] and [10], we instead adopt the ℓ_1 norm as a surrogate approximation to the ℓ_0 norm:

$$\|\mathbf{w}\|_1 = \sum_{i=0}^{N-1} |w_i|. \quad (14)$$

Using the regularization penalty $f_n(\mathbf{w}) = \|\mathbf{w}\|_1$ in regularized LMS (9), we obtain

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n + \mu_n e_n \mathbf{x}_n - \rho_n \text{sgn} \hat{\mathbf{w}}_n, \quad (15)$$

where the component-wise $\text{sgn}(\cdot)$ function is defined as

$$\text{sgn}(x) = \begin{cases} x/|x| & x \neq 0 \\ 0 & x = 0 \end{cases}. \quad (16)$$

Equation (15) yields the ZA-LMS introduced in [12]. The regularization parameter ρ_n can be calculated by (11) for LMS and by (12) for NLMS, where $f_n(\hat{\mathbf{w}}_n) = \|\hat{\mathbf{w}}_n\|_1$ and η_n is an estimate of the true $\|\mathbf{w}\|_1$.

An alternative approach to approximating the ℓ_0 norm is to consider the following function [12], [15], [19]:

$$\|\mathbf{w}\|_0 \simeq \sum_{i=0}^{N-1} \frac{1}{|w_i| + \delta} \cdot |w_i|, \quad (17)$$

where δ is a sufficiently small positive real number. Interpreting (17) as a weighted ℓ_1 approximation, we propose the regularization function $f_n(\mathbf{w})$

$$f_n(\mathbf{w}) = \sum_{i=0}^{N-1} \beta_{n,i} \cdot |w_i|, \quad (18)$$

and

$$\beta_{n,i} = \frac{1}{|\hat{w}_{n,i}| + \delta}, \quad (19)$$

where $\hat{w}_{n,i}$ is the i -th coefficient of $\hat{\mathbf{w}}_n$ defined in (1). Using (18) in (9) yields

$$\hat{w}_{n+1,i} = \hat{w}_{n,i} + \mu_n e_n x_{n-i} - \rho_n \beta_{n,i} \text{sgn} \hat{w}_{n,i}, \quad (20)$$

which is a component-wise update of the RZA-LMS proposed in [12]. Again, ρ_n can be computed using (11) for LMS or (12) for NLMS, where η_n is an estimate of the true $\|\mathbf{w}\|_0$, i.e., the number of the non-zero coefficients.

B. Group-sparse LMS

In many practical applications, a sparse system often exhibits a grouping structure, i.e., coefficients in the same group are highly correlated and take on the values zero or non-zero as a group, as shown in Fig. 1 (b). The motivation for developing group-sparse LMS is to take advantage of such a structure.

We begin by employing the mixed $\ell_{1,2}$ norm for promoting group-sparsity, which was originally proposed in [18] and has been widely adopted for various structured sparse regression problems [20], [21]. The $\ell_{1,2}$ norm of a vector \mathbf{w} is defined as

$$\|\mathbf{w}\|_{1,2} = \sum_{j=1}^J \|\mathbf{w}_{I_j}\|_2, \quad (21)$$

where $\{I_j\}_{j=1}^J$ is a group partition of the whole index set $I = \{0, 1, \dots, N-1\}$:

$$\bigcup_{j=1}^J I_j = I, \quad I_j \cap I_{j'} = \emptyset \text{ when } j \neq j', \quad (22)$$

and \mathbf{w}_{I_j} is a sub-vector of \mathbf{w} indexed by I_j . The $\ell_{1,2}$ norm is a mixed norm: it encourages correlation among coefficients inside each group via the ℓ_2 norm and promotes sparsity across those groups using the ℓ_1 norm. $\|\mathbf{w}\|_{1,2}$ is convex in \mathbf{w} and reduces to $\|\mathbf{w}\|_1$ when each group contains only one coefficient, i.e.,

$$|I_1| = |I_2| = \dots = |I_J| = 1, \quad (23)$$

where $|\cdot|$ denotes the cardinality of a set. Employing $f_n(\mathbf{w}) = \|\mathbf{w}\|_{1,2}$, the $\ell_{1,2}$ regularized LMS, which we refer to as GZA-LMS, is

$$\hat{\mathbf{w}}_{n+1,I_j} = \hat{\mathbf{w}}_{n,I_j} + \mu_n e_n \mathbf{x}_{I_j} - \rho_n \frac{\hat{\mathbf{w}}_{n,I_j}}{\|\hat{\mathbf{w}}_{n,I_j}\|_2 + \delta}, \quad j = 1, \dots, J, \quad (24)$$

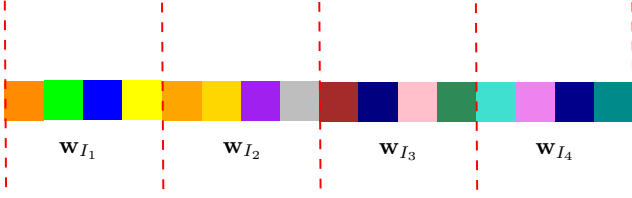


Fig. 2. A toy example illustrating the $\ell_{1,2}$ norm of a 16×1 coefficient vector \mathbf{w} : $\|\mathbf{w}\|_{1,2} = \sum_{j=1}^4 \|\mathbf{w}_{I_j}\|_2$.

and δ is a sufficiently small number ensuring a non-zero denominator. To the best of our knowledge this is the first time that the $\ell_{1,2}$ norm has been proposed for the LMS adaptive filters.

To further promote group selection we consider the following weighted $\ell_{1,2}$ regularization as a group-wise generalization of (18):

$$f_n(\mathbf{w}) = \sum_{j=1}^J \beta_{n,j} \|\mathbf{w}_{I_j}\|_2, \quad (25)$$

where $\beta_{n,j}$ is a re-weighting parameter defined by

$$\beta_{n,j} = \frac{1}{\|\hat{\mathbf{w}}_{n,I_j}\|_2 + \delta}, \quad (26)$$

and the corresponding regularized LMS update is then

$$\hat{\mathbf{w}}_{n+1,I_j} = \hat{\mathbf{w}}_{n,I_j} + \mu_n e_n \mathbf{x}_{I_j} - \rho_n \beta_{n,j} \frac{\hat{\mathbf{w}}_{n,I_j}}{\|\hat{\mathbf{w}}_{n,I_j}\|_2 + \delta}, \quad j = 1, \dots, J, \quad (27)$$

which is referred to as GRZA-LMS.

As both the $\ell_{1,2}$ norm and the weighted $\ell_{1,2}$ norm are convex, Theorem 1 applies under the assumption of white input signals and ρ_n can be calculated by (11) or (12). The parameter η_n can be chosen as an estimate of the true $\|\mathbf{w}\|_{1,2}$ for GZA-LMS (24), or the number of non-zero groups of \mathbf{w} for GRZA-LMS (27).

Finally, we note that GZA-LMS and GRZA-LMS reduce to ZA-LMS and RZA-LMS, respectively, if each group contains only one element.

C. Choosing regularization parameter for correlated input

Theorem 1 gives a closed form expression for ρ_n and (11) or (12) is applicable for any convex $f_n(\mathbf{w})$. However, the dominance over conventional LMS is only guaranteed when the input signal is white. Here we develop an alternative formula to determine ρ_n that applies to correlated input signals for sparse and group-sparse LMS, *i.e.*, (15), (20), (24) and (27).

We begin by considering the weighted $\ell_{1,2}$ regularization (25) and the corresponding GRZA-LMS update (27). Indeed, the other three algorithms, *i.e.*, (24), (20) and (15), can be treated as special cases of (27). For general wide-sense stationary (WSS) input signals, the regularization parameter ρ_n of (27) can be selected according the following theorem.

Theorem 2. Assume $\{x_n\}$ and $\{v_n\}$ are WSS stochastic processes which are mutually independent. Let $\hat{\mathbf{w}}_n$ and $\hat{\mathbf{w}}'_n$

be filter coefficients updated by (27) and (5) with the same μ_n , respectively. Then,

$$E \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|_2^2 \leq E \|\hat{\mathbf{w}}'_{n+1} - \mathbf{w}\|_2^2 \quad (28)$$

if $\hat{\mathbf{w}}_n = \hat{\mathbf{w}}'_n$ and $\rho_n \in [0, 2\rho_n^*]$, \mathbf{w} is the true coefficient vector and ρ_n^* is

$$\rho_n^* = \max \left\{ \frac{f_n(\hat{\mathbf{w}}_n) - \eta_n - \mu_n r_n}{\|\partial f_n(\hat{\mathbf{w}}_n)\|_2^2}, 0 \right\}, \quad (29)$$

where $f_n(\hat{\mathbf{w}}_n)$ is determined by (25), η_n is an upper bound of $f_n(\mathbf{w})$ and

$$r_n = \hat{\mathbf{w}}_n^T \mathbf{x}_n \cdot \mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n) + \eta_n \cdot \max_j \left\{ \frac{\|\mathbf{x}_{I_j}\|_2}{\beta_{n,j}} \right\} \cdot |\mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n)|. \quad (30)$$

The proof of Theorem 2 can be found in the Appendix. We make the following remarks.

Remark 4. Theorem 2 is derived from the general form (27) and can be directly specialized to (24), (20) and (15). Specifically,

- GZA-LMS (24) can be obtained by assigning $\beta_{n,j} = 1$;
- RZA-LMS (20) can be obtained when $|I_j| = 1, j = 1, \dots, J$;
- ZA-LMS (15) can be obtained when both $|I_j| = 1, j = 1, \dots, J$ and $\beta_{n,j} = 1$.

Remark 5. Theorem 2 is valid for any WSS input signals. However, the dominance result in (28) is weaker than that in Theorem 1, as it requires $\hat{\mathbf{w}}_n = \hat{\mathbf{w}}'_n$ at each iteration.

Remark 6. Eq. (29) can be applied to both LMS and NLMS, depending on if μ_n are deterministic functions of \mathbf{x}_n as specified in (6). This is different from Theorem 1 where we have separate expressions for LMS and NLMS.

Remark 7. ρ_n^* in (29) is non-zero only if $f_n(\hat{\mathbf{w}}_n)$ is greater than $\eta_n + \mu_n r_n$ (rather than η_n as presented in Theorem 1). This may yield a more conservative performance.

IV. NUMERICAL SIMULATIONS

In this section we demonstrate our proposed sparse LMS algorithms by numerical simulations. Multiple experiments are designed to evaluate their performances over a wide range of conditions.

A. Identifying a general sparse system

Here we perform evaluation of the proposed filters for general sparse system identification, as illustrated in Fig. 1 (a). There are 100 coefficients in the time varying system and only five of them are non-zero. The five non-zero coefficients are assigned to random locations and their values are also randomly drawn from a standard Gaussian distribution. The resultant true coefficient vector is plotted in Fig. 3.

1) *White input signals:* Initially we simulate white Gaussian input signal $\{x_n\}$ with zero mean and unit variance. The measurement noise $\{v_n\}$ is an independent Gaussian random process of zero mean and variance $\sigma_v^2 = 0.1$. For ease of parameter selection, we implement NLMS-type filters in our simulation. Three filters (NLMS, ZA-NLMS and RZA-NLMS) are implemented and their common step-size μ_n is set via (6)

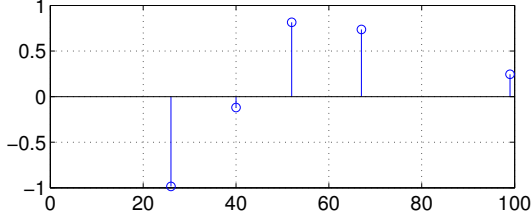


Fig. 3. The general sparse system used for simulations.

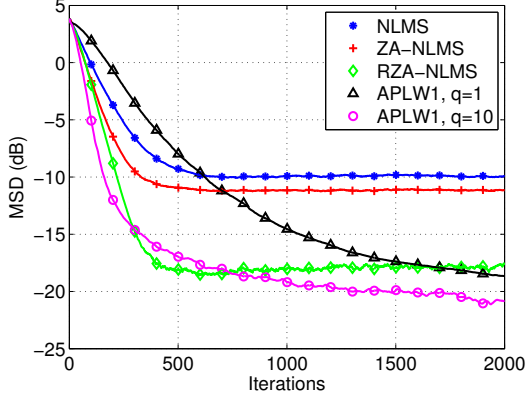


Fig. 4. White input signals: performance comparison for different filters.

with $\alpha_n = 1$. The regularization parameter ρ_n is computed using (12), where η_n is set to $\eta_n = \|\mathbf{w}\|_1$ (i.e., the true value) for ZA-NLMS and $\eta_n = 5$ for RZA-NLMS. For comparison we also implement a recently proposed sparse adaptive filter, referred to as APWL1 [15], which sequentially projects the coefficient vector onto weighted ℓ_1 balls. We note that our simulation setting is identical to that used in [15] and thus we adopt the same tuning parameters for APWL1. In addition, the weights $\beta_{n,i}$ for RZA-NLMS is scheduled in the same manner as that in [15] for a fair comparison. The simulations are run 100 times and the average estimates of mean square deviation (MSD) are shown in Fig. 4.

It can be observed that ZA-NLMS improves upon NLMS in both convergence rate and steady-state behavior and RZA-NLMS does even better. The parameter q of APLW1 is the number of samples used in each iteration. One can see that RZA-NLMS outperforms APLW1 when $q = 1$, i.e., the case that APLW1 operates with the same memory storage as RZA-NLMS. With larger p APLW1 begins to perform better and exceeds RZA-NLMS when $q \geq 10$. However, there is a trade-off between the system complexity and filtering performance, as APWL1 requires $\mathcal{O}(qN)$ for memory storage and $\mathcal{O}(N \log_2 N + qN)$ for computation, in contrast to LMS-type methods which require only $\mathcal{O}(N)$ for both memory and computation.

Next, we investigate the sensitivity to η_n for ZA-NLMS and RZA-NLMS. The result shown in Fig. 5 indicates that ZA-NLMS is more sensitive to η_n than RZA-NLMS, which is highly robust to misspecified η_n .

Further analysis reveals that the projection based methods such APWL1 may exhibit unstable converging behaviors. Fig.

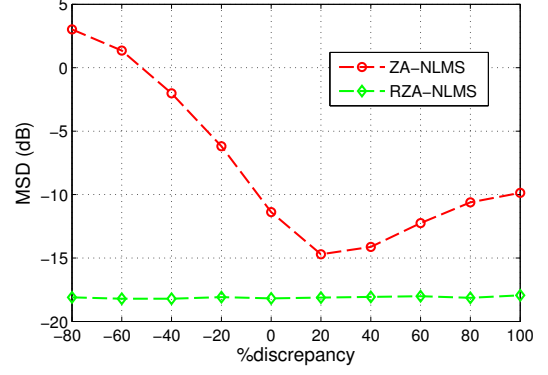


Fig. 5. Sensitivity of ZA-NLMS and RZA-NLMS to η_n : MSD for ZA-NLMS and RZA-NLMS at the 750th iteration for white input signals.

6 shows two independent trials of the simulation implemented in Fig. 4. It can be seen that there exist several local minima in APWL1. For example, Fig. 6 (b) seems to indicate that APWL1 ($q = 10$) converges at the 400th iteration with MSD $\simeq -12$ dB, yet its MSD actually reaches values as low as -25 dB at the 900th iteration. This slow convergence phenomenon is due to the fact that the weighted ℓ_1 ball is determined in an online fashion and the projection operator is sensitive to mis-specifications of the convex set. In the contrast, our regularized LMS uses sub-gradient rather than projection to pursue sparsity, translating into improved convergence.

2) *Correlated input signals*: Next, we evaluate the filtering performance using correlated input signals. We generate the sequence $\{x_n\}$ as an AR(1) process

$$x_n = 0.8x_{n-1} + u_n, \quad (31)$$

which is then normalized to unit variance, where $\{u_n\}$ is a Gaussian i.i.d. process. The measurement system is the same as before and the variance of the noise is also $\sigma_v^2 = 0.1$.

We compare our RZA-NLMS with APWL1 ($q = 10$) and standard NLMS is also included as a benchmark. All the filter parameters are set to the same values as that in the previous simulation, except we employ both (12) and (29) to calculate ρ_n in RZA-NLMS. The simulations are run 100 times and the average MSD curves are plotted in Fig. 7. While Theorem 1 is derived based on white input assumptions, using (12) to determine ρ_n achieves an empirically better performance compared to using (29) – whose use guarantees dominance but yields a conservative result. This confirms our conjecture in Remark 7. We also observe a severe performance degradation of APWL1 for correlated input signals. Fig. 8 draws two independent trials in this simulation. The phenomenon described in Fig. 6 becomes more frequent when the input signal is correlated, which drags down the average performance of APWL1 significantly. Finally, we note that the filtering performance of a group sparse system (e.g., Fig. 1 (b)) may be very different from that of a general sparse system. This will be investigated in Section IV-B.

3) *Tracking performance*: Finally, we study the tracking performance of the proposed filters. The time-varying system is initialized using the same parameters as used to generate

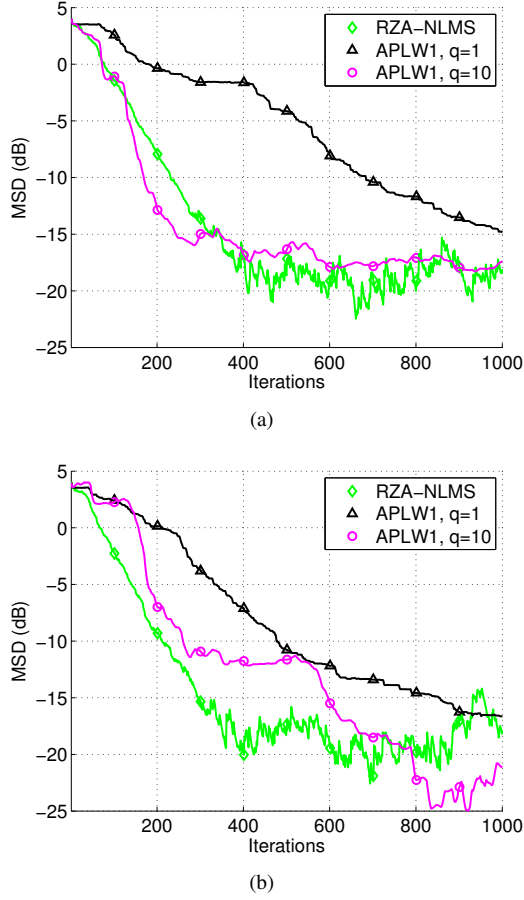


Fig. 6. Two different trials of RZA-NLMS and APWL1 for white input signals. APWL1 exhibits unstable convergence.

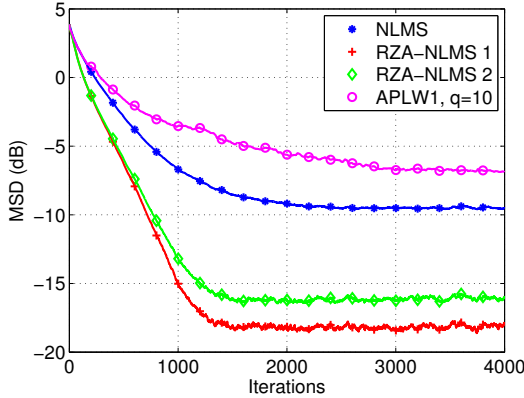


Fig. 7. Correlated input signals: performance comparison for different filters, where RZA-NLMS 1 and RZA-NLMS 2 use (12) and (29) to determine ρ_n , respectively.

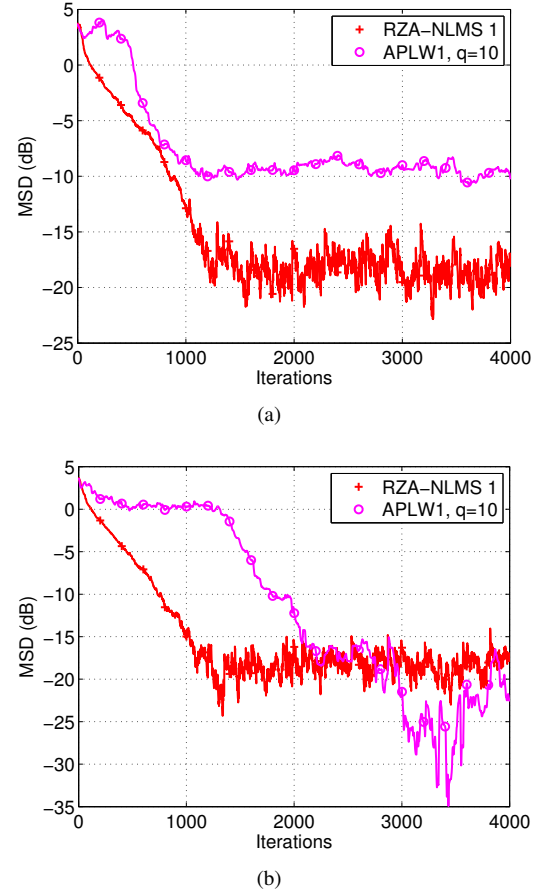


Fig. 8. Two different trials of RZA-NLMS and APWL1 for correlated input signals.

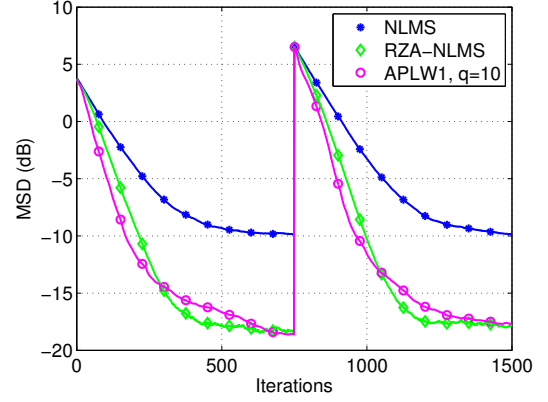


Fig. 9. Comparison of tracking performances when the input signal is white.

Fig. 3. At the 750th iteration the system encounters a sudden change, where all the active coefficients are left-shifted for 10 taps. We use white input signals to excite the unknown system and all the filter parameters are set in an identical manner to Section IV-A1. The simulation is repeated 100 times and the averaged result is shown in Fig. 9. It can be observed that both RZA-NLMS and APWL1 ($q = 10$) achieve better tracking performance than the conventional NLMS.

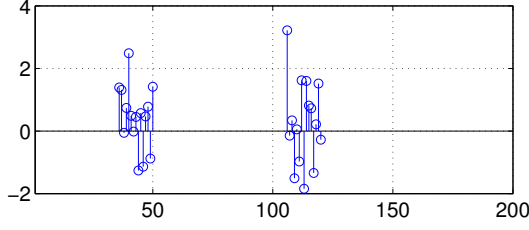


Fig. 10. The group-sparse system used for simulations. There are two active blocks; each of them contains 15 non-zero coefficients.

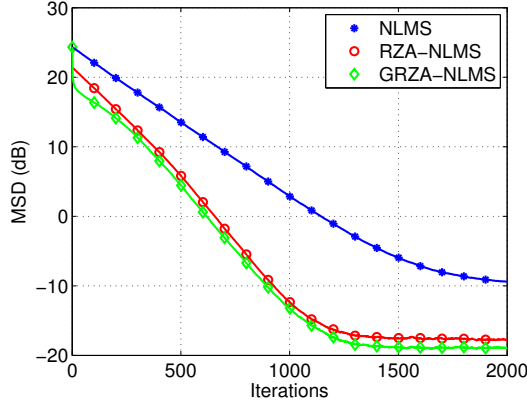


Fig. 11. MSD comparison for the group-sparse system for white input signals.

B. Identifying a group-sparse system

Here we test performance of the group-sparse LMS filters developed in Section III-B. The unknown system contains 200 coefficients that are distributed into two groups. The locations of the two groups are randomly selected, which start from the 36th tap and the 107th tap, respectively. Both of the two groups contain 15 coefficients and their values are randomly drawn from a standard Gaussian distribution. Fig. 10 shows the response of the true system.

The input signal $\{x_n\}$ is initially set to an i.i.d. Gaussian process and the variance of observation noise is $\sigma_v^2 = 0.1$. Three filters, GRZA-NLMS, RZA-NLMS and NLMS, are implemented, where the performance of NLMS is treated as a benchmark. In GRZA-NLMS, we divide the 200 coefficients equally into 20 groups, where each of them contains 10 coefficients. The step size μ_n of the three filters are all set according to (6) with $\alpha_n = 1$. We use (12) to calculate ρ_n , where η_n is set to 30 (the number of non-zero coefficients) for RZA-NLMS and 2 (the number of non-zero blocks) for GRZA-NLMS, respectively. We repeat the simulation 200 times and the averaged MSD is shown in Fig. 11. It can be seen that GRZA-NLMS and RZA-NLMS outperform the standard NLMS for 10 dB in the steady-state MSD, while GRZA-NLMS only improves upon RZA-NLMS, but only marginally. This is partially due to the fact that in the white input scenario each coefficient is updated in an independent manner.

We next consider the case of correlated input signals, where $\{x_n\}$ is generated by (31) and then normalized to have unit variance. The parameters for all the filters are set to the same values as in the white input example and the averaged MSD

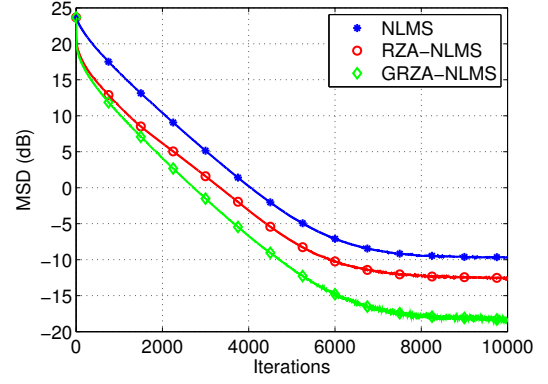


Fig. 12. MSD comparison for the group-sparse system for correlated input signals.

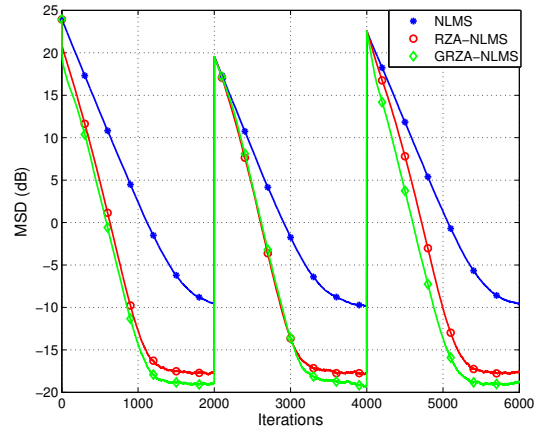


Fig. 13. Tracking performance comparison for the group-sparse system for white input signals.

curves are plotted in Fig. 12. In the contrast to the white input example, here RZA-NLMS slightly outperforms NLMS but there is a significant improvement of GRZA-NLMS over RZA-NLMS. This demonstrates the power of promoting group-sparsity especially when the input signal is correlated.

Finally, we evaluate the tracking performance of the adaptive filters. We use white signals as the system input and initialize the time-varying system using that in Fig. 10. At the 2000th iteration, the system response is right-shifted for 50 taps, while the values of coefficients inside each block are unaltered. We then keep the block locations and reset the values of non-zero coefficients randomly at the 4000th iteration. From Fig. 13 we observe that the tracking rate of RZA-NLMS and GRZA-NLMS are comparable to each other when the system changes across blocks, and GRZA-NLMS shows a better tracking performance than RZA-NLMS when the system response changes only inside its active groups.

V. CONCLUSION

In this paper we proposed a general class of LMS-type filters regularized by convex sparsifying penalties. We derived closed-form expressions for choosing the regularization parameter that guarantees provable dominance over conventional

LMS filters. We applied the proposed regularized LMS filters to sparse and group-sparse system identification and demonstrated their performances using numerical simulations.

Our regularized LMS filter is derived from the LMS framework and inherits its simplicity, low computational cost and low memory requirements, and robustness to parameter mismatch. It is likely that the convergence rate and steady-state performance can be improved by extension to second-order methods, such as RLS and Kalman filters. Efficient extensions of our results for sparse/group-sparse RLS filters are a worthy topic of future study.

VI. APPENDIX

A. Proof of Theorem 1

We prove Theorem 1 for LMS, *i.e.*, the case that μ_n are constants. NLMS, where μ_n is determined by (6), can be derived in a similar manner.

According to (9),

$$\begin{aligned} \hat{\mathbf{w}}_{n+1} - \mathbf{w} &= (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)(\hat{\mathbf{w}}_n - \mathbf{w}) - \rho_n \partial f_n(\hat{\mathbf{w}}_n) + \mu_n v_n \mathbf{x}_n. \end{aligned} \quad (32)$$

Noting that $\hat{\mathbf{w}}_n$, \mathbf{x}_n and v_n are mutually independent, we have

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 | \hat{\mathbf{w}}_n \} &= (\hat{\mathbf{w}}_n - \mathbf{w})^T E \left\{ (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 \right\} (\hat{\mathbf{w}}_n - \mathbf{w}) + \mu_n^2 \sigma_v^2 E \{ \|\mathbf{x}_n\|^2 \} \\ &\quad + 2\rho_n (\mathbf{w} - \hat{\mathbf{w}}_n)^T E \left\{ \mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T \right\} \partial f_n(\hat{\mathbf{w}}_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|^2. \end{aligned} \quad (33)$$

As $\{x_n\}$ is a Gaussian i.i.d. process, \mathbf{x}_n is a Gaussian random vector with mean zero and covariance $\sigma_x^2 \mathbf{I}$. Thus,

$$E \left\{ (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 \right\} = (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) \mathbf{I}, \quad (34)$$

$$E \left\{ \mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T \right\} = (1 - \sigma_x^2 \mu_n) \mathbf{I}, \quad (35)$$

and

$$E \{ \|\mathbf{x}_n\|^2 \} = N\sigma_x^2. \quad (36)$$

Substituting (34), (35) and (36) into (33), we have

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 | \hat{\mathbf{w}}_n \} &= (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) \|\hat{\mathbf{w}}_n - \mathbf{w}\|^2 + N\mu_n^2 \sigma_x^2 \sigma_v^2 \\ &\quad + 2\rho_n (1 - \sigma_x^2 \mu_n) (\mathbf{w} - \hat{\mathbf{w}}_n)^T \partial f_n(\hat{\mathbf{w}}_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|^2. \end{aligned} \quad (37)$$

As $f_n(\cdot)$ is a convex function, by the definition of sub-gradient, we have

$$(\mathbf{w} - \hat{\mathbf{w}}_n)^T \partial f_n(\hat{\mathbf{w}}_n) \leq f_n(\mathbf{w}) - f_n(\hat{\mathbf{w}}_n) \leq \eta_n - f_n(\hat{\mathbf{w}}_n). \quad (38)$$

Therefore,

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 | \hat{\mathbf{w}}_n \} &\leq (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) \|\hat{\mathbf{w}}_n - \mathbf{w}\|^2 + N\mu_n^2 \sigma_x^2 \sigma_v^2 \\ &\quad - 2\rho_n (1 - \sigma_x^2 \mu_n) (f_n(\hat{\mathbf{w}}_n) - \eta_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|^2. \end{aligned} \quad (39)$$

Define

$$C(\rho_n) = -2\rho_n (1 - \sigma_x^2 \mu_n) (f_n(\hat{\mathbf{w}}_n) - \eta_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|^2, \quad (40)$$

and take expectation on both sides of (39) with respect to $\hat{\mathbf{w}}_n$ to obtain

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 \} &\leq (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) E \{ \|\hat{\mathbf{w}}_n - \mathbf{w}\|^2 \} + N\mu_n^2 \sigma_x^2 \sigma_v^2 \\ &\quad + E \{ C(\rho_n) \}. \end{aligned} \quad (41)$$

It is easy to check that $C(\rho_n) \leq 0$ if $\rho_n \in [0, 2\rho_n^*]$, where ρ_n^* is defined in (11). Therefore,

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 \} &\leq (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) E \{ \|\hat{\mathbf{w}}_n - \mathbf{w}\|^2 \} + N\mu_n^2 \sigma_x^2 \sigma_v^2 \end{aligned} \quad (42)$$

if $\rho_n \in [0, 2\rho_n^*]$. For the standard LMS, there is

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}'_{n+1} - \mathbf{w}\|^2 \} &= (1 - 2\sigma_x^2 \mu_n + N\sigma_x^4 \mu_n^2) E \{ \|\hat{\mathbf{w}}'_n - \mathbf{w}\|^2 \} + N\mu_n^2 \sigma_x^2 \sigma_v^2. \end{aligned} \quad (43)$$

Therefore, under the condition that $E \{ \|\hat{\mathbf{w}}_0 - \mathbf{w}\|^2 \} = E \{ \|\hat{\mathbf{w}}'_0 - \mathbf{w}\|^2 \}$, (10) can be obtained from (42) and (43) using a simple induction argument.

B. Proof of Theorem 2

We start our proof from (32) and calculate the following conditional MSD:

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 | \hat{\mathbf{w}}_n, \mathbf{x}_n \} &= (\hat{\mathbf{w}}_n - \mathbf{w})^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 (\hat{\mathbf{w}}_n - \mathbf{w}) + \mu_n^2 \sigma_v^2 \|\mathbf{x}_n\|^2 + D(\rho_n), \end{aligned} \quad (44)$$

where

$$D(\rho_n) = 2\rho_n (\mathbf{w} - \hat{\mathbf{w}}_n)^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T) \partial f_n(\hat{\mathbf{w}}_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|^2. \quad (45)$$

For the cross term $2\rho_n (\mathbf{w} - \hat{\mathbf{w}}_n)^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T) \partial f_n(\hat{\mathbf{w}}_n)$ we have

$$\begin{aligned} 2\rho_n (\mathbf{w} - \hat{\mathbf{w}}_n)^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T) \partial f_n(\hat{\mathbf{w}}_n) &= 2\rho_n (\mathbf{w} - \hat{\mathbf{w}}_n)^T \partial f_n(\hat{\mathbf{w}}_n) + 2\rho_n \mu_n \hat{\mathbf{w}}_n^T \mathbf{x}_n \cdot \mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n) \\ &\quad - 2\rho_n \mu_n \mathbf{w}^T \mathbf{x}_n \cdot \mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n) \\ &\leq 2\rho_n (\eta_n - f_n(\hat{\mathbf{w}}_n)) + 2\rho_n \mu_n \hat{\mathbf{w}}_n^T \mathbf{x}_n \cdot \mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n) \\ &\quad + 2\rho_n \mu_n |\mathbf{w}^T \mathbf{x}_n| \cdot |\mathbf{x}_n^T \partial f_n(\hat{\mathbf{w}}_n)|. \end{aligned} \quad (46)$$

We now establish upper-bounds for $|\mathbf{w}^T \mathbf{x}_n|$. Indeed,

$$\begin{aligned} |\mathbf{w}^T \mathbf{x}_n| &= \left| \sum_{j=1}^J \mathbf{w}_{I_j}^T \mathbf{x}_{n, I_j} \right| \\ &\leq \sum_{j=1}^J \left| \beta_{n,j} \mathbf{w}_{I_j}^T \frac{1}{\beta_{n,j}} \mathbf{x}_{n, I_j} \right| \\ &\leq \sum_{j=1}^J \beta_{n,j} \|\mathbf{w}_{I_j}\|_2 \frac{\|\mathbf{x}_{n, I_j}\|_2}{\beta_{n,j}} \\ &\leq \left\{ \sum_{j=1}^J \beta_{n,j} \|\mathbf{w}_{I_j}\|_2 \right\} \max_j \frac{\|\mathbf{x}_{n, I_j}\|_2}{\beta_{n,j}} \\ &= f_n(\mathbf{w}_n) \max_j \frac{\|\mathbf{x}_{n, I_j}\|_2}{\beta_{n,j}} \leq \eta_n \max_j \frac{\|\mathbf{x}_{n, I_j}\|_2}{\beta_{n,j}}. \end{aligned} \quad (47)$$

Substituting (46) and (47) into (45) we obtain that

$$D(\rho_n) \leq -2\rho_n(f_n(\hat{\mathbf{w}}_n) - \eta_n - \mu_n r_n) + \rho_n^2 \|\partial f_n(\hat{\mathbf{w}}_n)\|_2^2, \quad (48)$$

where r_n is defined in (30). Note that $D(\rho_n) \leq 0$ if $\rho_n \in [0, 2\rho_n^*]$ (ρ_n^* is defined in (29)). There is

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 | \hat{\mathbf{w}}_n, \mathbf{x}_n \} \\ \leq (\hat{\mathbf{w}}_n - \mathbf{w})^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 (\hat{\mathbf{w}}_n - \mathbf{w}) + \mu_n^2 \sigma_v^2 \|\mathbf{x}_n\|^2, \end{aligned} \quad (49)$$

if $\rho_n \in [0, 2\rho_n^*]$. Therefore,

$$\begin{aligned} E \{ \|\hat{\mathbf{w}}_{n+1} - \mathbf{w}\|^2 \} \\ \leq E \{ (\hat{\mathbf{w}}_n - \mathbf{w})^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 (\hat{\mathbf{w}}_n - \mathbf{w}) \} \\ + \mu_n^2 \sigma_v^2 E \{ \|\mathbf{x}_n\|^2 \} \\ = E \{ (\hat{\mathbf{w}}'_n - \mathbf{w})^T (\mathbf{I} - \mu_n \mathbf{x}_n \mathbf{x}_n^T)^2 (\hat{\mathbf{w}}'_n - \mathbf{w}) \} \\ + \mu_n^2 \sigma_v^2 E \{ \|\mathbf{x}_n\|^2 \} \\ = E \{ \|\hat{\mathbf{w}}'_{n+1} - \mathbf{w}\|^2 \}, \end{aligned} \quad (50)$$

which proves Theorem 2.

REFERENCES

- [1] B. Widrow and S.D. Stearns, *Adaptive Signal Processing*, New Jersey: Prentice Hall, 1985.
- [2] S. Kawamura and M. Hatori, "A TAP selection algorithm for adaptive filters," in *Proceedings of ICASSP*, 1986, vol. 11, pp. 2979–2982.
- [3] J. Homer, I. Mareels, R.R. Bitmead, B. Wahlberg, and A. Gustafsson, "LMS estimation via structural detection," *IEEE Trans. on Signal Processing*, vol. 46, pp. 2651–2663, October 1998.
- [4] Y. Li, Y. Gu, and K. Tang, "Parallel NLMS filters with stochastic active taps and step-sizes for sparse system identification," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, IEEE, 2006, vol. 3.
- [5] D.M. Etter, "Identification of sparse impulse response systems using an adaptive delay filter," in *Proceedings of ICASSP*, 1985, pp. 1169–1172.
- [6] M. Godavarti and A. O. Hero, "Partial update LMS algorithms," *IEEE Trans. on Signal Processing*, vol. 53, pp. 2382–2399, 2005.
- [7] S.L. Gay, "An efficient, fast converging adaptive filter for network echocancellation," in *Proceedings of Asilomar*, 1998, vol. 1, pp. 394–398.
- [8] D.L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 508–518, 2000.
- [9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B.*, vol. 58, pp. 267–288, 1996.
- [10] E. Candès, "Compressive sampling," *Int. Congress of Mathematics*, vol. 3, pp. 1433–1452, 2006.
- [11] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, March 2007.
- [12] Y. Chen, Y. Gu, and A.O. Hero, "Sparse LMS for system identification," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, pp. 3125–3128.
- [13] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 8, pp. 4013–4025, 2010.
- [14] D. Angelosante, J.A. Bazerque, and G.B. Giannakis, "Online Adaptive Estimation of Sparse Signals: Where RLS Meets the ℓ_1 -Norm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 7, pp. 3436–3447, 2010.
- [15] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online Sparse System Identification and Signal Reconstruction using Projections onto Weighted ℓ_1 Balls," *Arxiv preprint arXiv:1004.3040*, 2010.
- [16] Y. Gu, J. Jin, and S. Mei, " ℓ_0 Norm Constraint LMS Algorithm for Sparse System Identification," *IEEE Signal Processing Letters*, vol. 16, pp. 774–777, 2009.
- [17] W.F. Schreiber, "Advanced television systems for terrestrial broadcasting: Some problems and some proposed solutions," *Proceedings of the IEEE*, vol. 83, no. 6, pp. 958–981, 1995.
- [18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

- [19] E.J. Candes, M.B. Wakin, and S.P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.
- [20] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.
- [21] F.R. Bach, "Consistency of the group Lasso and multiple kernel learning," *The Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.